Maximum likelihood estimation of structure parameters from high resolution electron microscopy images. Part I: A theoretical framework

A.J. den Dekker\textsuperscript{a,*}, S. Van Aert\textsuperscript{b}, A. van den Bos\textsuperscript{c}, D. Van Dyck\textsuperscript{b}

\textsuperscript{a}Delft Center for Systems and Control, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands

\textsuperscript{b}University of Antwerp, Department of Physics, Groenenborgerlaan 171, 2020 Antwerp, Belgium

\textsuperscript{c}Delft University of Technology, Faculty of Applied Sciences, Lorentzweg 1, 2628 CJ Delft, The Netherlands

Received 4 October 2004; received in revised form 23 February 2005; accepted 2 March 2005

Abstract

This paper is the first part of a two-part paper on maximum likelihood (ML) estimation of structure parameters from electron microscopy images. In principle, electron microscopy allows structure determination with a precision that is orders of magnitude better than the resolution of the microscope. This requires, however, a quantitative, model-based method. In our opinion, the ML method is the most appropriate one since it has optimal statistical properties. This paper aims to provide microscopists with the necessary tools to apply this method so as to determine structure parameters as precisely as possible. It reviews the theoretical framework, including model assessment, the derivation of the ML estimator of the parameters, the limits to precision and the construction of confidence regions and intervals for ML parameter estimates. In a companion paper [Van Aert et al., Ultramicroscopy, this issue, 2005], a practical example will be worked out.

© 2005 Elsevier B.V. All rights reserved.

PACS: 61.16.B; 42.30.W; 43.50.+y; 07.05.Fb

Keywords: High-resolution transmission electron microscopy (HRTEM); Electron microscope design and characterization; Data processing/image processing

1. Introduction

The increasing need for precise determination of the atomic arrangement of nonperiodic structures in materials design and control of nanostructures explains the growing interest into quantitative
electron microscopy. Unlike qualitative methods, which are based on visual interpretation of images, quantitative electron microscopy allows the extraction of local structure information at the subangstrom level. The starting point of quantitative electron microscopy is the notion that one is not so much interested in the electron microscopy images as such, but rather in the structure information of the object under study. Therefore, images are to be considered as data planes, from which structure parameters have to be determined as precisely as possible. The key to successful quantitative electron microscopy image analysis is the availability of a pertinent parametric model of the images. The model includes all ingredients needed to perform a computer simulation of the image. It is parametric in the quantities of interest, such as the locations of the atom columns, which are usually unknown beforehand and have to be estimated from the experiment. The authors admit that it is probably still premature to assume the availability of such a model. It is true that the electron–matter interactions and the image formation process are nowadays sufficiently well understood to allow a calculation of electron microscopy images and several commercial software packages for high-resolution transmission electron microscopy (HRTEM) image simulation are available [2,3]. The thus simulated images show close qualitative agreement with the experimental images in the sense that image details are described adequately. However, the contrast in experimental images is commonly two or three times lower than in simulated images. Obviously, this so-called contrast problem [4] hampers today’s quantitative structure determination. Researchers currently investigate various physical phenomena which are not yet taken into account in the model and which may explain the contrast problem [4–8]. Until the contrast problem is solved, it may also be circumvented by using simplified, empirical models that describe the experimental contrast adequately, although some model components may lack a solid physical basis. For example, King et al. try to avoid the contrast problem by adding a constant parameter, in the form of a background, to the models currently used in image simulation packages [9]. Another example of an image model that is simplified, but adequate for the purpose of the measurement procedure, which is the measurement of certain structure parameters, is given in part two of this paper.

For the moment, let us assume the availability of a parametric model that accurately describes the images. Methods to test the validity of the model will be presented in this paper. Each experimental image consists of a set of pixels values. In this paper, these pixel values will be called the observations. Due to the inevitable presence of noise, experimental images (i.e., sets of observations) made under the same conditions will nevertheless fluctuate randomly, that is, they will differ from experiment to experiment. The usual way to describe this behavior is to model the observations as stochastic variables [10]. By definition, a stochastic variable is characterized by its probability density function (distribution), while a set of stochastic variables has a joint probability density function. It is the expectation (mean value) of each observation that is described by the parametric model mentioned above. The parameterized probability density function can be used for two purposes [10]. First, it may be used to define the concept of Fisher information and to compute the Cramér–Rao lower bound (CRLB), which is a lower bound on the variance of any unbiased estimator of the parameters. Second, from the probability density function of the observations the maximum likelihood (ML) estimator of the parameters may be derived. This estimator actually achieves the CRLB asymptotically, that is, for an infinite number of observations. Therefore, the ML estimator is asymptotically most precise. For this and other reasons, the ML estimator is very important and often used in practice. In order to evaluate the level of confidence to be attached to the particular values of the ML estimates obtained, confidence regions and intervals (often referred to as ‘error bars’) associated with these estimates are required. Ideally, these confidence regions and intervals are obtained by repeating the same experiment over and over again (under stationary conditions), followed by an evaluation of the statistical distribution of the estimates obtained. However, this scenario is often not very realistic in practice. Alternative ways to construct
confidence regions and intervals have been proposed in the literature and will be reviewed in this paper.

This paper is Part I of a two-part paper. In the accompanying paper, Part II [1], an experimental study of an aluminium crystal is presented so as to show the practical applicability of the theory described in Part I.

The organization of this paper is as follows. Section 2 discusses the derivation of a parametric statistical model of the observations. In Section 3, the concept of Fisher information and its relation to the precision with which model parameters can be estimated will be introduced. Section 4 describes the derivation of the ML estimator. Model assessment is discussed in Section 5. Section 6 reviews methods to construct confidence regions and intervals for the parameter estimates obtained by applying the ML method. In Section 7, conclusions are drawn.

2. Parametric statistical model of the observations

The usual way to describe the fluctuating behavior of images in the presence of noise is modelling the image pixel values, which will be called the observations, as stochastic variables. Consider a set of observations \( \{w_n, n = 1, \ldots, N\} \) is available and let \( w \) define the \( N \times 1 \) vector of observations:

\[
w = (w_1 \ldots w_N)^T,
\]

where the superscript \( T \) denotes transposition. If the observations are counting results, they are nonnegative integers and the probability density function defines the probability of occurrence of each of these integer outcomes. If the observations are continuous, the probability density function describes the probability of occurrence of an observation on a particular interval [10]. A set of observations is defined by its joint probability density function. The expectation value (mean value) of each observation, \( \mathbb{E}[w_n] \), is defined by its probability density function. In this paper, it is assumed that these expectation values \( \mathbb{E}[w_n] \) can be described by a functional model \( f_n(\tau) \) which is parametric in the quantities \( \tau \) to be measured (such as the locations of atom columns):

\[
\mathbb{E}[w_n] = f_n(\tau)
\]

with \( \tau = (\tau_1 \ldots \tau_R)^T \) the \( R \times 1 \) vector of unknown parameters to be measured. The availability of such a model makes it possible to parametrize the probability density function of the observations, which is of vital importance to precise quantitative structure determination. An example is given below.

**Example 1.** Let us consider statistically independent Poisson distributed observations \( w_n, n = 1, \ldots, N \). For example, these observations could be electron counting results detected with a CCD camera with a quantum efficiency equal to 1. The probability that the observation \( w_n \) is equal to \( \omega_n \) is then equal to:

\[
\frac{(\lambda_n)^{\omega_n}}{\omega_n!} \exp(-\lambda_n),
\]

where the parameter \( \lambda_n \) is equal to the expectation \( \mathbb{E}[w_n] \). Furthermore, since the observations are independent, the probability \( p(\omega) \) of a set of observations \( \omega = (\omega_1 \ldots \omega_N)^T \) is the product of all probabilities described by Eq. (3):

\[
p(\omega) = \prod_{n=1}^{N} \frac{(\lambda_n)^{\omega_n}}{\omega_n!} \exp(-\lambda_n).
\]

Next, suppose that the expectations \( \lambda_n \) can be described by a functional model \( f_n(\tau) \):

\[
\mathbb{E}[w_n] \equiv \lambda_n = f_n(\tau)
\]

with \( \tau = (\tau_1 \ldots \tau_R)^T \) the \( R \times 1 \) vector of unknown parameters to be measured. For instance, \( f_n(\tau) \) may be given by

\[
f_n(\tau) = \zeta + \frac{\eta}{2\pi\rho^2} \exp\left(-\frac{(x_n - \beta_x)^2 + (y_n - \beta_y)^2}{2\rho^2}\right)
\]

with \( \tau = (\zeta, \eta, \beta_x, \beta_y, \rho)^T \). This model may, for example, represent a Gaussian shaped atom column superposed on a background. In this model, \( x_n \) and \( y_n \) then represent the assumedly exactly known measurement points, whereas \( \zeta, \eta, \beta_x, \beta_y \) and \( \rho \) denote the unknown parameters representing the background, the amplitude, position coordinates and width of the atom column,
respectively. The latter parameters have to be estimated from the observations. The quantities $w_n - \lambda_n$ are the so-called nonsystematic errors in the observations. If the expectation model (5) is correct, their expectations are equal to zero.

Example 1 shows that the expectation of the observations is an accurate description of what experimenters often call the model underlying the observations. In quantitative electron microscopy, this model corresponds to the model that is fitted to the images in a refinement procedure. This model includes all ingredients needed to perform a computer simulation of the image. Furthermore, substitution of $\lambda_n$ as described by Eq. (6) in Eq. (4) shows how the probability density function depends on the unknown parameters to be measured. Probability density functions thus parameterized can be used for two purposes. First, they may be used to define the concept of Fisher information and to compute the CRLB, which is a lower bound on the variance of any unbiased estimator of the parameters. This will be the subject of Section 3. Second, from the probability density function of the observations the ML estimator of the parameters may be derived, which will be the subject of Section 4.

Before proceeding, the notation that is used in what follows should be explained. Let $x$ and $y$ be $R \times 1$ and $L \times 1$ vectors, respectively, let $f(x)$ be a scalar function of the elements of $x$ and let $g(x)$ be an $L \times 1$ vector function of $x$. Then:

(a) the $1 \times R$ vector $\partial f(x)/\partial x$ is defined by its $r$th element $\partial f(x)/\partial x_r$;

(b) the $R \times R$ matrix $\partial^2 f(x)/\partial x^2$ is defined by its $(r, s)$th element $\partial^2 f(x)/\partial x_r \partial x_s$;

(c) the $L \times R$ matrix $\partial g(x)/\partial x$ is defined by its $(l, r)$th element $\partial g_l/\partial x_r$;

(d) the $L \times R$ matrix $\partial^2 g(x)/\partial x^2$ is defined by its $(l, r)$th element $\partial^2 g_l/\partial x_r \partial x_s$.

3. Fisher information and attainable precision

Let $p(w; \tau)$ be the joint probability density function of a set of observations $\{w_n, n = 1, \ldots, N\}$. Example 1 makes clear how the dependence of $p(w; \tau)$ on the parameters $\tau$ to be measured can be established. This dependence can now be used to define the so-called Fisher score [11]. The Fisher score vector is defined as the $1 \times R$ vector:

$$s(\tau) = \frac{\partial \ln p(w; \tau)}{\partial \tau}. \quad (7)$$

It can be shown that the expectation of the Fisher score vector is equal to the null vector:

$$E[s(\tau)] = E\left[\frac{\partial \ln p(w; \tau)}{\partial \tau}\right] = 0. \quad (8)$$

Notice that this result applies to any allowable parameterization of the probability density function [10]. As a consequence, the $R \times R$ covariance matrix of the Fisher score is described by

$$F = E \left[\frac{\partial \ln p(w; \tau)}{\partial \tau} \frac{\partial \ln p(w; \tau)}{\partial \tau}^T\right]. \quad (9)$$

Furthermore, by the multi-variate central limit theorem, the Fisher score vector will tend to be multivariate normally distributed [12]. It is not difficult to show that $F$ may alternatively be written as

$$F = -E \left[\frac{\partial^2 \ln p(w; \tau)}{\partial \tau^2}\right]. \quad (10)$$

This covariance matrix is called the Fisher information matrix. The Fisher information is a measure of the physical fluctuations in the observations. In a sense, the Fisher information specifies the quality of the measurements. It expresses the ‘inability to know’ a measured quantity [13]. Indeed, use of the concept of Fisher information allows one to determine the highest precision, that is, the lowest variance, with which a parameter can be estimated unbiasedly. Suppose that $\hat{\tau}$ is any unbiased estimator of $\tau$, that is, $E[\hat{\tau}] = \tau$. Then it can be shown that under general conditions the covariance matrix $\text{cov}(\hat{\tau})$ of $\hat{\tau}$ satisfies [14]

$$\text{cov}(\hat{\tau}) \geq F^{-1}. \quad (11)$$

This inequality expresses that the difference between the left-hand and right-hand member is positive semi-definite. A property of a positive semi-definite matrix is that its diagonal elements cannot be negative. This means that the diagonal
elements of \( \text{cov}(\hat{\tau}) \), that is, the variances of \( \hat{\tau}_1, \ldots, \hat{\tau}_R \) are larger than or equal to the corresponding diagonal elements of \( F^{-1} \). In this sense \( F^{-1} \) represents a lower bound to the variances of all unbiased \( \hat{\tau} \). The matrix \( F^{-1} \) is called the CRLB on the variance of \( \tau \). The derivation of the CRLB is illustrated in Examples 2 and 3.

Notice that the CRLB is not related to a particular estimation method. It depends on the statistical properties of the observations, the measurement points, and in most cases, the hypothetical true values of the parameters. This dependence on the true values looks, at first sight, as a serious impediment to the practical use of the bound. However, the expressions for the bound provide the experimenter with quantitative insight in what precision might be achieved from the available observations. In addition, it provides insight in the sensitivity of the precision to the parameter values [15].

In what follows, the \( N \times 1 \) expectation model vector function \( f(\tau) \) is defined as

\[
f(\tau) = (f_1(\tau) \cdots f_N(\tau))^T \tag{12}
\]

and the \( N \times R \) Jacobian matrix \( J \) of the expectation model evaluated at \( \tau \) is defined as

\[
J(\tau) = \frac{\partial f(\tau)}{\partial \tau}. \tag{13}
\]

Furthermore, let \( q(\omega) = \ln p(\omega) \).

**Example 2 (Poisson distributed observations).** The joint probability density function of a set of independent Poisson distributed observations \( \{w_n, n = 1, \ldots, N\} \) is given by

\[
p(\omega) = \prod_n \frac{(\lambda_n \omega_n)^{\omega_n}}{\omega_n!} \exp(-\lambda_n) \tag{14}
\]

with \( \omega = (\omega_1 \cdots \omega_N)^T \). The logarithm of Eq. (14) is then described by

\[
q(\omega) = \sum_n -\lambda_n + \omega_n \ln \lambda_n - \ln \omega_n! \tag{15}
\]

Suppose that the expectation values \( \mathbb{E}[w_n] \equiv \lambda_n \) can be described by a functional model which is parametric in the parameters to be measured: \( \mathbb{E}[w_n] = f_n(\tau) \), with \( \tau = (\tau_1 \ldots \tau_R)^T \) the vector of parameters. Then, it follows from Eqs. (9), (13) and (15) that the Fisher information matrix can be described as

\[
F = J^T(\tau) \text{diag}(f_1^{-1}(\tau) \cdots f_N^{-1}(\tau)) J(\tau) \tag{16}
\]

with \( \text{diag}(f_1^{-1}(\tau) \cdots f_N^{-1}(\tau)) \) a diagonal matrix of order \( N \) with the elements \( f_1^{-1}(\tau) \cdots f_N^{-1}(\tau) \) on the diagonal. Then it can be shown that the elements of the Fisher information matrix (16) can be described as

\[
F_{il} = \sum_n \frac{1}{f_n(\tau)} \frac{\partial f_n(\tau)}{\partial \tau_i} \frac{\partial f_n(\tau)}{\partial \tau_l}, \quad r = 1, \ldots, R, \quad l = 1, \ldots, R. \tag{17}
\]

The CRLB can now be found by taking the inverse of the Fisher information matrix.

Notice that for a scalar valued parameter \( \tau \), it follows from Eqs. (11) and (17) that

\[
\text{var}(\hat{\tau}) \geq \frac{1}{\sum_n \frac{1}{f_n(\tau)} \left( \frac{\partial f_n(\tau)}{\partial \tau} \right)^2}. \tag{18}
\]

**Example 3 (normally distributed observations).** Let us consider a set of normally distributed observations \( \{w_n, n = 1, \ldots, N\} \) of which the expectations \( \mathbb{E}[w] \) can be described by the functional model \( f(\tau) \) parametric in the parameters \( \tau \) to be measured. Then, the joint probability density function of these observations is given by

\[
p(\omega; \tau) = \frac{1}{(2\pi)^{N/2} |det W|^{1/2}} \times \exp \left[ -\frac{1}{2} (\omega - f(\tau))^T W^{-1}(\omega - f(\tau)) \right] \tag{19}
\]

with \( W \) the \( N \times N \) covariance matrix of the observations with as \( (n_1, n_2) \)th element \( \text{cov}(w_{n_1}, w_{n_2}) \). Furthermore, \( \text{det} \ W \) and \( W^{-1} \) are the determinant and the inverse of \( W \), respectively. The logarithm of \( p(\omega; \tau) \) is equal to

\[
q(\omega; \tau) = -\frac{N}{2} \ln 2\pi - \frac{1}{2} \ln \text{det} \ W - \frac{1}{2} (\omega - f(\tau))^T W^{-1}(\omega - f(\tau)). \tag{20}
\]
Then, it follows from Eqs. (9), (13) and (20) that the Fisher information matrix is given by

\[ F = J^T(\tau)W^{-1}J(\tau). \]  

(21)

Consequently, the CRLB is given by

\[ F^{-1} = (J^T(\tau)W^{-1}J(\tau))^{-1}. \]  

(22)

For uncorrelated observations, the covariance matrix \( W \) is given by \( W = \text{diag}(\sigma_{w_1}^2, \ldots, \sigma_{w_N}^2) \), with \( \sigma_{w_n}^2 \) the variance of the \( n \)th observation. If, in addition, \( \sigma_{w_n}^2 = \sigma^2 \) for all \( n \), the joint probability density function (19), its logarithm (20), and the Fisher information matrix (21) simplify to

\[ p(\omega; \tau) = \prod_n \frac{1}{(2\pi)^{N/2}} \exp \left[ -\frac{(\omega_n - f_n(\tau))^2}{2\sigma^2} \right], \]  

(23)

\[ q(\omega; \tau) = -\frac{N}{2} \ln 2\pi - \frac{N}{2} \ln \sigma - \frac{1}{2\sigma^2} \sum_n (\omega_n - f_n(\tau))^2, \]  

(24)

and

\[ F = \frac{1}{\sigma^2} J^T(\tau)J(\tau), \]  

(25)

respectively.

Then, it follows from Eqs. (11) and (25) that for a scalar valued parameter \( \tau \)

\[ \text{var}(\hat{\tau}) \geq \frac{\sigma^2}{\sum_n \left( \frac{\partial f_n(\tau)}{\partial \tau} \right)^2}. \]  

(26)

Up to now, \( \sigma^2 \) is supposed to be known. If it is not known, it can be estimated simultaneously with the parameters \( \tau \). Let \( \theta = (\tau^T \sigma^2)^T \) denote the new \((R+1) \times 1\) parameter vector. Then, it follows from Eqs. (10) and (24) that the \((R+1) \times (R+1)\) Fisher information matrix with respect to \( \theta \) is given by

\[ F_\theta = \begin{pmatrix} F & 0 \\ 0 & \frac{N}{2\sigma^2} \end{pmatrix}, \]  

(27)

where \( F \) is the \( R \times R \) Fisher information matrix given by (25). The CRLB is thus given by

\[ F_\theta^{-1} = \begin{pmatrix} F^{-1} & 0 \\ 0 & \frac{2\sigma^2}{N} \end{pmatrix}. \]  

(28)

The CRLB can be extended to include unbiased estimators of vectors of functions of the parameters instead of the parameters as such. Let \( \rho(\tau) = (\rho_1(\tau) \ldots \rho_L(\tau))^T \) be such a vector and let \( \hat{\rho} = (\hat{\rho}_1 \ldots \hat{\rho}_L)^T \) be an unbiased estimator of \( \rho(\tau) \). Then it can be shown that

\[ \text{cov}(\hat{\rho}) \geq \frac{\partial \rho}{\partial \tau}^T \cdot F_\theta^{-1} \cdot \left( \frac{\partial \rho}{\partial \tau} \right). \]  

(29)

Experimental design. Another important purpose for which the expressions for the CRLB can be used is the optimization of the experimental design. Experimental design can be defined as the selection of free variables in an experiment to improve the precision of the measured parameters. By calculating the CRLB, the experimenter gets an impression if for a given experimental design the precision attainable is sufficient for the purposes concerned. If not, the experimental design has to be changed. If this is not possible, it has to be concluded that the observations are not suitable for the purposes of the measurement procedure. In this way, the experimental design can be optimized so as to attain the highest precision [15]. Indeed, it has been shown recently how optimizing the design of quantitative electron microscopy experiments may enhance the precision with which structure parameters can be estimated [16–21].

As demonstrated above, every set of observations contains a certain amount of Fisher information. It is up to the experimenter to extract this information from the observations. For this purpose, different estimators may be used. Generally, different estimators have different precisions. The most precise estimator is the one that attains the CRLB, i.e., the one that extracts all Fisher information. It is known that there exists an estimator that achieves the CRLB at least asymptotically, that is, for an increasing number of observations. This estimator is the ML estimator. The ML estimator is easily derived and has a
number of favorable statistical properties [22]. It will be introduced in Section 4.

4. ML estimation

4.1. The derivation of the ML estimator

The derivation of the maximum likelihood estimator, which requires the probability density function of the observations and its dependence on the unknown parameters to be known, consists of the following steps [10]:

- First, the available observations \( w = (w_1 \ldots w_N)^T \) are substituted for the corresponding variables \( \omega = (\omega_1 \ldots \omega_N)^T \) in the probability density function of the observations. Since the observations are numbers, the resulting expression only depends on the elements of \( \tau = (\tau_1 \ldots \tau_R)^T \).
- Next, the elements of \( \tau = (\tau_1 \ldots \tau_R)^T \), which are the hypothetical true parameters, are considered to be variables. To express this, they are replaced by \( t = (t_1 \ldots t_R)^T \). The resulting function \( p(t) \) is called the likelihood function of the parameters \( t \) for the observations \( w \).
- Finally, the ML estimates \( \hat{\tau}_{\text{ML}} \) of the parameters \( \tau \) are computed. These are defined as the values of the elements of \( t \) that maximize \( p(t) \):

\[
\hat{\tau}_{\text{ML}} = \arg \max_t p(t) \quad (30)
\]

or, equivalently,

\[
\hat{\tau}_{\text{ML}} = \arg \max_t q(t) \quad (31)
\]

in which \( q(t) \) is called the log-likelihood function.

The construction of the ML estimator from Poisson and normally distributed observations is illustrated in the Examples 4–6.

Example 4 (Poisson distributed observations).

Suppose that a set of independent Poisson distributed observations \( \{w_n : n = 1, \ldots, N\} \) is available. Their parametrized joint probability density function is described by substituting Eq. (5) in Eq. (4). Substituting the available observations \( w = (w_1 \ldots w_N)^T \) for the corresponding variables \( \omega = (\omega_1 \ldots \omega_N)^T \) and replacing the true parameters \( \tau = (\tau_1 \ldots \tau_R)^T \) by \( t = (t_1 \ldots t_R)^T \) then yields the likelihood function:

\[
p(t) = \prod_n \frac{(f_n(t))^{w_n}}{w_n^!} \exp(-f_n(t)). \quad (32)
\]

Taking the logarithm yields:

\[
q(t) = \sum_n \left[ -f_n(t) + w_n \ln f_n(t) - \ln w_n! \right] \quad (33)
\]

The ML estimates \( \hat{\tau}_{\text{ML}} \) of the parameters \( \tau \) are then given by

\[
\hat{\tau}_{\text{ML}} = \arg \max_t \sum_n \left[ -f_n(t) + w_n \ln f_n(t) \right]. \quad (34)
\]

Example 5 (normally distributed observations with known covariance matrix).

Suppose that a set of normally distributed observations \( \{w_n : n = 1, \ldots, N\} \) is available and that their parametrized joint probability density function can be described by Eq. (19) with known covariance matrix \( W \). Substituting the available observations \( w = (w_1 \ldots w_N)^T \) for the corresponding independent variables \( \omega = (\omega_1 \ldots \omega_N)^T \) in Eq. (19) and replacing the true parameters \( \tau = (\tau_1 \ldots \tau_R)^T \) by \( t = (t_1 \ldots t_R)^T \) then yields the likelihood function:

\[
p(t) = \frac{1}{(2\pi)^{W/2}(\det W)^{1/2}} \times \exp \left[ -\frac{1}{2} (w - f(t))^T W^{-1} (w - f(t)) \right]. \quad (35)
\]

Taking the logarithm of (35) then yields the log-likelihood function

\[
q(t) = -\frac{N}{2} \ln 2\pi - \frac{1}{2} \ln \det W
\]

\[
- \frac{1}{2} (w - f(t))^T W^{-1} (w - f(t)). \quad (36)
\]

The ML estimates \( \hat{\tau}_{\text{ML}} \) of the parameters \( \tau \) are then given by

\[
\hat{\tau}_{\text{ML}} = \arg \min_t (w - f(t))^T W^{-1} (w - f(t)). \quad (37)
\]

For uncorrelated observations, with \( W = \text{diag}(\sigma_{w_1}^2 \ldots \sigma_{w_N}^2) \), Eq. (37) becomes:

\[
\hat{\tau}_{\text{ML}} = \arg \min_t \sum_n \frac{(w_n - f_n(t))/\sigma_{w_n}^2}{\sigma_n^2} \quad (38)
\]
where \( \sigma_w^2 \) is the variance of \( \omega_n \). If, in addition, 
\( \sigma_w^2 = \sigma^2 \) for all \( n \):

\[
\hat{\sigma}_{\text{ML}} = \arg \min_t \sum_n [w_n - f_n(t)]^2. \tag{39}
\]

From the above, it is clear that for normally distributed observations the ML estimator corresponds to the weighted least-squares estimator. The weight factors used depend on the covariance matrix of the observations. For uncorrelated observations with equal variance, they are uniform. In that case, knowledge of the noise variance is not required to derive the ML estimator of the parameters \( \tau \). Nevertheless, the noise variance may be estimated simultaneously with the parameters \( \tau \) using the ML method, as will be illustrated in Example 6.

**Example 6** (independent normally distributed observations with unknown but equal variance). Consider independent normally distributed observations \( \{w_n, n = 1, \ldots, N\} \) of which the parametrized joint probability density function is described by Eq. (19), with \( W = \text{diag}(\sigma^2 \ldots \sigma^2) \). Suppose that the variance \( \sigma^2 \) of the observations is unknown and that one wants to estimate \( \sigma^2 \) simultaneously with the elements of the parameter vector \( \tau \). As in Example 3, we may define a new \( (R + 1) \times 1 \) parameter vector \( \theta \) as \( \theta = (\tau^T \sigma^2)^T \).

Substituting the available observations \( w = (w_1 \ldots w_N)^T \) for the corresponding independent variables \( \omega = (\omega_1 \ldots \omega_N)^T \) and replacing the true parameters \( \tau \) and \( \sigma^2 \) by \( \hat{\tau} \) and \( \hat{s}^2 \), respectively, then yields the likelihood function:

\[
p(t, s^2) = \prod_n \frac{1}{(2\pi)^{1/2} s} \exp \left[ -\frac{(w_n - f_n(t))^2}{2s^2} \right]. \tag{40}
\]

The ML estimates \( \hat{\tau}_{\text{ML}} \) and \( \hat{\sigma}_{\text{ML}}^2 \) are then given by

\[
(\hat{\tau}_{\text{ML}}, \hat{\sigma}_{\text{ML}}^2) = \arg \max_{t,s^2} \left[ -\frac{N}{2} \ln 2\pi - \frac{N}{2} \ln s^2 - \frac{1}{2s^2} \sum_n (w_n - f_n(t))^2 \right]. \tag{41}
\]

In this case it is convenient to use the following steps of maximization [23]. The first step is to find the unique value \( \psi(t, w) \) of \( s^2 \) which maximizes \( p(t, s^2) \) with respect to \( s^2 \), with \( t \) regarded as a constant. It can be shown that

\[
\psi(t, w) = \frac{1}{N} (w - f(t))^T (w - f(t)) = \frac{1}{N} \sum_n (w_n - f_n(t))^2. \tag{42}
\]

The second step consists of finding that value \( \hat{\tau} = \bar{\tau}(w) \) of \( t \) which maximizes

\[
M(t) \equiv q(t, \psi(t, w)) = -\frac{N}{2} \left( \ln 2\pi + \ln \sum_n (w_n - f_n(t))^2 - \ln N + 1 \right). \tag{43}
\]

where \( M(t) \) is called the concentrated log-likelihood function [23]. It can be proven that the thus obtained values \( \hat{\tau} \) and \( \psi(\hat{\tau}, w) \) are equal to the ML estimators of \( \tau \) and \( \sigma^2 \), respectively. Notice that \( M(t) \) is maximized when \( \sum_n (w_n - f_n(t))^2 \) is minimized, that is, when \( t \) equals (39) (the least-squares estimate).

**4.2. Asymptotic statistical properties of the ML estimator**

The ML estimator has a number of favorable statistical properties [22]. First, it can be shown that this estimator achieves the CRLB asymptotically, that is, for an infinite number of observations. Therefore, it is asymptotically most precise (or, asymptotically efficient). Second, it can be shown that the ML estimator is consistent, which means that it converges to the true value of the parameter in a statistically well defined way if the number of observations increases. Third, the ML estimator is asymptotically normally distributed, with a mean equal to the true value of the parameter and a covariance matrix equal to the CRLB. If these asymptotic properties also apply to a finite or even small number of observations can often only be assessed by estimating from artificial, simulated observations.

**4.3. The need for a good starting structure**

The relative ease with which the likelihood function can be derived, does not mean that its
maximum can be found easily. Finding the ML estimate corresponds to finding the global maximum in a $R$-dimensional parameter space, where $R$ is the number of parameters to be estimated. In quantitative structure determination, each possible combination of the $R$ parameters corresponds directly to an atomic structure. It is represented by a point in a $R$-dimensional parameter space. The search for the maximum of the likelihood function is usually an iterative numerical procedure. In electron microscopy applications, the dimension of the parameter space is usually very high. Consequently, it is quite possible that the optimization procedure ends up at a local maximum instead of at the global maximum of the likelihood function, so that the wrong structure model is suggested, which introduces a systematic error (bias). To solve this dimensionality problem, that is, to find a pathway to the maximum in the parameter space, good starting values for the parameters are required. In other words, the structure has to be resolved. This corresponds to X-ray crystallography, where one first has to resolve the structure and afterwards one has to refine the structure. Resolving the structure is not trivial. It is known that details in HRTEM images do not necessarily correspond to features in the atomic structure. This is not only due to the unavoidable presence of noise, but also to the dynamic scattering of the electrons on their way through the object, and the image formation in the electron microscope which both have a blurring effect. As a consequence, the structure information of the object may be strongly delocalized which makes it very difficult to find good starting values for the structure parameters. However, it has been shown that starting values for the parameters, and hence a good starting structure of the object, may be found by using so-called direct methods. Direct methods in a sense invert the imaging process and the dynamic scattering process using some prior knowledge which is generally valid, irrespective of the (unknown) structure parameters of the object. The starting structure obtained with such a direct method is called a pseudoinverse [24]. A pseudoinverse can be obtained in different ways. The imaging process can be inverted by using direct methods for reconstruction of the so-called exit wave, i.e., the complex wave function at the exit face of the object. These methods include exit wave reconstruction

- from a series of images taken from the same area at different defocus values [25–27];
- from an electron holographic image [28];
- from a series of images recorded with different illuminating beam tilts [29];

Since the exit wave is more directly related to the object structure than the original image, exit wave reconstruction methods may provide appropriate starting values for the structure parameters. In general, however, there exists no direct one-to-one relationship between the projected structure and the exit wave, due to the complex nature of the dynamic scattering of the electrons on their way through the object. This applies particularly to thick objects containing heavy elements and to crystalline objects aligned along a low-index zone axis with respect to the incident electron beam [30]. Therefore, even if one has the disposal of exit waves, it may still be hard to find appropriate starting values that guarantee convergence to the global optimum of the likelihood function. To solve this problem, one may try to invert the dynamic scattering process as well. For this purpose, several direct methods that extract structure information from the reconstructed exit waves have been proposed [30–40]. Indeed, successive application of direct methods inverting the imaging and dynamic scattering process may yield an approximate structure model that can be used as a starting structure for ML estimation from the original images.

5. Model assessment

As mentioned above, ML estimation requires the probability density function of the observations and its dependence on the unknown parameters to be known. It has been shown in this paper that this dependence is often established by the availability of a parametric model that describes the expectations of the observations. Substitution of this so-called expectation model in
the expression for the probability density function provides a parameterized probability density function from which the ML estimator can be derived. Obviously, such an estimation procedure is valid if and only if the expectation model is an accurate description of the true expectations \( \theta \). Therefore, it is important to test the validity of the expectation model before attaching confidence to the parameter estimates obtained. If the model is inadequate, it must be modified and the analysis continued until a satisfactory result is obtained [41]. Many statistical model assessment methods are based on specific distributional assumptions. In particular, the assumption that the deviations of the observations from their expectations (i.e., the noise contributions) are independent, zero mean normally distributed with common variance, is common in classical statistical tests. The normality assumption is often justified by appealing to the central limit theorem, which states that the resultant of independent, zero mean normally distributed expectations (i.e., the noise contributions) are valid if and only if the expectation model is valid. In particular, the assumption that methods are based on specific distributional assumptions. In Section 5.1, the Likelihood Ratio (LR) test will be discussed, which is a test that is applicable to any probability density function of the observations. In Section 5.2, some model assessment methods based on the normality assumption will be reviewed.

5.1. The LR test

A test that is applicable to any probability density function of the observations and does not need replications (obtained from repeating the same experiment), is the so-called LR test, which can be described as follows. Suppose that we have a set of observations \( \{w_n, n = 1, \ldots, N\} \) and that we want to test the null hypothesis \( H_0 \) against the alternative hypothesis \( H_1 \), with

- \( H_0 \): The expectations of the observations are described by the functional model \( f_n(\tau) \)
- \( H_1 \): The expectations of the observations are not described by the functional model \( f_n(\tau) \)

where \( \tau = (\tau_1, \ldots, \tau_N)^T \) and \( R \leq N \). The null hypothesis \( H_0 \) can be restated as a set of restrictions on the parameters used to formulate the likelihood function of the data. This can be understood as follows. Consider the general model \( g_n(\theta) = \theta_n \), with \( \theta = (\theta_1, \ldots, \theta_N)^T \). It can be shown that \( \{f_n(\tau)\} \subset \{g_n(\theta)\} \). In words, the models \( g_n(\theta) \) and \( f_n(\tau) \) are nested, that is, the model \( g_n(\theta) \) includes the expectation model \( f_n(\tau) \), of which we wish to test the validity, as a special case. Indeed, the model \( f_n(\tau) \) can be obtained from \( g_n(\theta) \) by imposing \( \theta_n = f_n(\tau) \), which comes down to putting \( N - R \) independent equality constraints on the parameters \( \theta \). This effectively reduces the number of unknown parameters from \( N \) to \( R \). Testing the validity of the expectation model \( f_n(\tau) \) then corresponds to testing the hypothesis consisting of these (generally non-linear) equality constraints.

Let \( q(\theta) \) be the log-likelihood function which is parametric in the parameters \( \theta \). Furthermore, let \( \theta_{\text{ML}} \) be the unrestricted ML estimate of \( \theta \), which is obtained by unconstrained maximizing of \( q(\theta) \) with respect to \( \theta \). In addition, let \( \theta_{\text{RL}} \) be the restricted ML estimate of \( \theta \), which is obtained by maximizing \( q(\theta) \) under the above mentioned equality constraints on \( \theta \). Since these constraints effectively transform the model \( g_n(\theta) \) into the model \( f_n(\tau) \), it can easily be shown that \( q(\theta_{\text{ML}}) \) equals \( q(\theta_{\text{RL}}) \).

**Test statistic.** Then it can be shown that, when \( H_0 \) is true, the statistic [23]

\[
\text{LR} = 2[q(\hat{\theta}_{\text{ML}}) - q(\hat{\theta}_{\text{RL}})]
\]

(44)
is approximately \( \chi^2_{N-R} \) distributed, i.e., chi-square distributed with \( N - R \) degrees of freedom. This approximation, which becomes exact asymptotically, is usually adequate, even for a finite number of observations. Furthermore, notice that LR will always be positive, since the maximum value of the likelihood function for restricted parameters, will always be smaller than or equal to the maximum value of the likelihood function for unrestricted parameters. The likelihood ratio test uses LR as a test statistic.

**Decision rule.** The range of LR values for which the hypothesis \( H_0 \) is accepted is given by the interval

\[
(\chi^2_{N-R,\alpha/2}, \chi^2_{N-R,1-\alpha/2})
\]

(45)
where \( \gamma \) corresponds to the so-called test size chosen by the experimenter and \( \chi^2_{N-R/2} \) and \( \chi^2_{N-R,1-\gamma/2} \) represent the \((\gamma/2)\) and \((1-\gamma/2)\) quantile of a chi-square distribution with \( N-R \) degrees of freedom, respectively [42]. The size of the test \( \gamma \) is also defined as the significance level, corresponds with the probability of rejecting \( H_0 \) if \( H_0 \) is true. Consequently, for a test with test size \( \gamma \), the probability of accepting \( H_0 \) if \( H_0 \) is true is \((1-\gamma)\). Furthermore, the \( q \) quantile of the distribution of a continuous random variable \( X \) is defined as the smallest number \( \xi \) satisfying \( S_X(\xi) = q \), with \( S_X(x) \) the cumulative distribution function of \( X \). Notice that alternative decision rules based on the test statistic (44) can be used. For example, one could choose to reject the hypothesis \( H_0 \) if the value of LR is larger than \( \chi^2_{N-R,1-\gamma} \). Example 7 illustrates the derivation of the LR that can be used as a test statistic to test the validity of a model assumed to describe the expectations of Poisson distributed observations.

**Example 7** (Poisson distributed observations). Suppose that a set of Poisson distributed observations \( w = (w_1 \ldots w_N)^T \) is available and that we wish to test the hypothesis that the expectations of these observations can be described by the functional model \( f_p(\tau) \), which is parametric in the elements of the parameter vector \( \tau = (\tau_1 \ldots \tau_R)^T \), with \( R \leq N \). As a more general model, we consider \( g_n(\theta) = \theta_n \), with \( \theta = (\theta_1 \ldots \theta_N)^T \). The model \( f_p(\tau) \) is a special case of the general model \( g_n(\theta) = \theta_n \). It can be constructed by putting \( N-R \) equality constraints (i.e., restrictions) on the parameters \( \theta \). The ML estimate of the parameters \( \theta \) of the unrestricted model \( g_n(\theta) \) is given by

\[
\hat{\theta}_{ML} = \arg \max_{\theta} \sum_n [-\theta_n + w_n \ln \theta_n] \tag{46}
\]

which yields \( \hat{\theta}_{ML} = w \). The log-likelihood function evaluated at the unrestricted ML estimate \( \hat{\theta}_{ML} \) is thus given by

\[
q(\hat{\theta}_{ML}) = q(w) = \sum_n [-w_n + w_n \ln w_n - \ln w_n!] \tag{47}
\]

The log-likelihood function evaluated at the restricted ML estimate \( \hat{\theta}_{ML} \) is given by

\[
q(\hat{\theta}_{ML}) = \sum_n [-\hat{\theta}_{ML,n} + w_n \ln[\hat{\theta}_{ML,n} - \ln w_n]]
\]

with \( \hat{\theta}_{ML,n} \) the \( n \)th element of \( \hat{\theta}_{ML} \). After reparameterization, Eq. (48) equals

\[
q(\tilde{\tau}_{ML}) = \sum_n [-f_p(\tilde{\tau}_{ML}) + w_n \ln f_p(\tilde{\tau}_{ML}) - \ln w_n!]
\]

with \( \tilde{\tau}_{ML} \) the ML estimate of the parameter vector \( \tau \). It follows from Eqs. (44), (47) and (49) that the likelihood ratio test statistic LR is given by

\[
LR = 2 \sum_n [-w_n + w_n \ln w_n - w_n \ln f_p(\tilde{\tau}_{ML}) + f_p(\tilde{\tau}_{ML})].
\]

Notice that since \( \lim_{x \to 0} x \ln x = 0 \), the term \( w_n \ln w_n \) reduces to zero for \( w_n = 0 \).

Similarly, the following example illustrates the derivation of the LR which can be used as a test statistic to test the validity of a model assumed to describe the expectations of normally distributed observations.

**Example 8** (normally distributed observations). Suppose that a set of normally distributed observations \( \{w_n, n = 1, \ldots, N\} \) is available and that we wish to test the assumption that the expectations of these observations can be described by the functional model \( f_p(\tau) \), which is parametric in the elements of the parameter vector \( \tau = (\tau_1 \ldots \tau_R)^T \), with \( R \leq N \). As a more general model, we consider \( g_n(\theta) = \theta_n \), with \( \theta = (\theta_1 \ldots \theta_N)^T \). The model \( f_p(\tau) \) is a special case of the general model \( g_n(\theta) = \theta_n \). It can be obtained by putting \( N-R \) restrictions on the parameters \( \theta \). The ML estimate of the parameters \( \theta \) of the unrestricted model \( g_n(\theta) \) is given by

\[
\hat{\theta}_{ML} = \arg \min_{\theta} (w - \theta)^T W^{-1}(w - \theta) \tag{51}
\]

which yields \( \hat{\theta}_{ML} = w \). The log-likelihood function evaluated at the unrestricted ML estimate \( \theta \) is then given by

\[
q(\hat{\theta}_{ML}) = -\frac{N}{2} \ln 2\pi - \frac{1}{2} \ln \text{det } W. \tag{52}
\]
The log-likelihood function evaluated at the restricted ML estimate \( \bar{\theta}_{\text{ML}} \) is given by

\[
q(\bar{\theta}_{\text{ML}}) = -\frac{N}{2} \ln 2\pi - \frac{1}{2} \ln \det W \\
- \frac{1}{2} (w - \bar{\theta}_{\text{ML}})^T W^{-1} (w - \bar{\theta}_{\text{ML}})
\]  

(53)

which, after reparameterization, equals

\[
q(\bar{\tau}_{\text{ML}}) = -\frac{N}{2} \ln 2\pi - \frac{1}{2} \ln \det W \\
- \frac{1}{2} (w - f(\bar{\tau}_{\text{ML}}))^T W^{-1} (w - f(\bar{\tau}_{\text{ML}}))
\]  

(54)

with \( \bar{\tau}_{\text{ML}} \) the ML estimate of the parameter vector \( \tau \). It follows from Eqs. (44), (52) and (54) that

\[
\text{LR} = (w - f(\bar{\tau}_{\text{ML}}))^T W^{-1} (w - f(\bar{\tau}_{\text{ML}})).
\]  

(55)

For uncorrelated observations the test statistic (55) becomes equal to

\[
\text{LR} = \sum_{n=1}^{N} \frac{(w_n - f_n(\bar{\tau}_{\text{ML}}))^2}{\sigma^2_{w_n}},
\]  

(56)

where \( \sigma^2_{w_n} \) is the variance of the \( n \)th observation. If, in addition, all observations have the same variance \( \sigma^2 \), the test statistic (55) becomes equal to

\[
\text{LR} = \sum_{n} \frac{(w_n - f_n(\bar{\tau}_{\text{ML}}))^2}{\sigma^2}.
\]  

(57)

Apart from the LR test, other test methods and associated test statistics have been developed, such as Wald’s test, and the Lagrange multiplier test. It has been reported, that under fairly general conditions, the statistics associated with the LR test, the Wald test and the Lagrange multiplier test are all asymptotically equivalent and asymptotically \( \chi^2_{N-R} \) distributed when \( H_0 \) is true [23].

5.2. Model assessment based on the assumption of normality

5.2.1. Residual plot

A simple but effective graphical method for checking the validity of the proposed expectation model is to make a plot of the residuals versus the measurement points (i.e., pixel positions). The residuals are defined as the deviations of the observations \( w_n \) from the expectation model \( f_n(\tau) \) evaluated at the ML estimate \( \bar{\tau}_{\text{ML}} \):

\[
e_n = w_n - f_n(\bar{\tau}_{\text{ML}}), \quad n = 1, \ldots, N.
\]  

(58)

Recall that it was shown in Section 4 that for independent, identically normally distributed noise contributions, the ML estimator equals the ordinary (i.e., uniformly weighted) least-squares estimator. Residual plots may reveal systematic (i.e., nonrandom) deviations of the model from the observations. Systematic deviations are deviations that cannot be due to the random (assumed to be normally distributed) noise contributions. The occurrence of systematic deviations tends to indicate an inadequacy of the expectation model, that is, the model is not able to describe the expectations of the observations since the functional dependence of these expectations on the parameters has not been incorporated correctly in the model. In this way, modelling errors, or equivalently, systematic errors, are introduced. Obviously, nonrandom behavior of the residuals may also indicate that the assumption of independent, normally distributed noise with constant variance is invalid. For example, an increase of the variability of the residuals with the observed image intensity level may indicate that the noise contributions are not identically normally distributed, but that their noise variance increases with the image intensity level. One may then modify the estimation procedure by using weighted (instead of uniformly weighted) least squares, or by applying the ML estimator for Poisson distributed observations (since the variance of a Poisson distributed observation equals its expectation). Moreover, tendencies for the residuals to stay positive or negative in certain areas of the image can reveal dependence or correlation of the noise (it follows from Eq. (19) that for normally distributed observations zero correlation implies independence and vice versa). If the noise disturbances are correlated, the ML estimator has to be modified so as to take account of correlation. Methods to reveal, quantify and account for (normally distributed) correlated noise distributions in least-squares estimation are discussed in Refs. [41,43].
5.2.2. Normal probability plot

An appropriate method to test the assumption that the noise contributions are independent, identically normally distributed (with unspecified mean and variance) is a so-called normal probability plot [44]. If the expectation model is correct and the assumption of normality is appropriate, a normal probability plot of the residuals should be a fairly straight line [41]. Departures from a straight line suggest inappropriateness of the proposed parametric statistical model of the observations in the sense that either the expectation model or the normality assumption is invalid. Such departures indicate that the shapes of the distribution of the residuals and the theoretical distribution do not match and may also suggest the nature of this mismatch [44]. Normal probability plots are also useful for revealing outliers. MATLAB’s statistics toolbox [45] provides a function, called normplot.m, for drawing normal probability plots.

5.2.3. Quantitative tests for normality

The normality of the residuals can also be tested quantitatively instead of graphically. For that purpose, the well-known Kolmogorov–Smirnov test [42], as well as the so-called Lilliefors test [46] can be used. The difference between the Kolmogorov–Smirnov test and the Lilliefors test is that the Kolmogorov–Smirnov assumes the mean and variance of the distribution to be known, whereas the Lilliefors test does not make this assumption. Instead, the mean and variance are estimated from the data.

5.2.3.1. The Kolmogorov–Smirnov test. Let us assume for the moment that it is known that the noise contributions have zero mean and known noise variance $\sigma^2$. Then the Kolmogorov–Smirnov test can be applied. The Kolmogorov–Smirnov test tests how well a given sample fits some completely specified distribution [42]. If it is used to decide if the residuals are independent, normally distributed with zero mean and variance $\sigma^2$, the null hypothesis $H_0$ and the alternative hypothesis $H_1$ are given by

$H_0$: the residuals are independent, normally distributed with zero mean and variance $\sigma^2$ and

$H_1$: the residuals are not independent, normally distributed with zero mean and variance $\sigma^2$.

If the null hypothesis is rejected, one has to conclude that the proposed expectation model is inadequate and/or that the assumption of zero mean identically normally distributed noise contributions is wrong. Anyway, a rejection of the null hypothesis makes clear that ML estimation based on the proposed expectation model and noise distribution would not be justified.

Test statistic. The Kolmogorov–Smirnov test statistic is defined as

$$T_{KS} = \sup_x |S_{en}(x) - S_0(x)|$$

(59)

with $S_{en}(x)$ the so-called empirical cumulative distribution function of the sample of $e_n$ values, which is defined by

$$S_{en}(x) = \frac{1}{N} \times \text{(number of residuals } e_n \text{ less than or equal to } x)$$

(60)

with $N$ the number of observations and $S_0(x)$ the completely specified cumulative distribution function associated with the normal distribution with zero mean and variance $\sigma^2$. In Eq. (59), $T_{KS}$ is a random variable that measures how far $S_{en}(x)$ deviates from $S_0(x)$. Its distribution under $H_0$ has been tabulated (see, e.g., Ref. [47]).

Decision rule. Given a user-specified test size $\gamma$, the hypothesis $H_0$ is rejected if $T_{KS}$ exceeds the $(1 - \gamma)$ quantile as given in a table of quantiles of the Kolgomorov–Smirnov test for normality as provided, for example, in Ref. [47]. MATLAB’s statistics toolbox [45] provides a function, called kstest.m, that performs a Kolmogorov–Smirnov test at any user-specified test size.

5.2.3.2. The Lilliefors test. Next, let us assume that the assumedly normally distributed noise contributions have unknown mean and variance. Then the Lilliefors test can be applied. In fact, the Lilliefors test is equal to the Kolmogorov–Smirnov test with estimated mean and variance. For the
Lilliefors test, the null hypothesis $H_0$ and the alternative hypothesis $H_1$ are given by

$H_0$: the residuals are independent and normally distributed, with unspecified mean and variance and

$H_1$: the residuals are not independent and normally distributed.

Again, if the null hypothesis is rejected, one has to conclude that the proposed expectation model is inadequate and/or that the assumption of independent, normally distributed noise contributions is wrong. A rejection of the null hypothesis thus makes clear that ML estimation based on the proposed expectation model and noise distribution would not be justified.

**Test statistic.** An expression for the Lilliefors test statistic, which is used to decide whether to accept or reject the null hypothesis, is obtained as follows. First, the sample mean and sample variance of the residuals $e_n$, which are defined by Eq. (58), are computed. The sample mean is defined by

$$\bar{e} = \frac{1}{N} \sum_{n=1}^{N} e_n$$

and the sample variance is defined by

$$s^2 = \frac{1}{N-1} \sum_{n=1}^{N} (e_n - \bar{e})^2.$$  \hspace{1cm} (62)

Next, the residuals are converted to the values $z_n = (e_n - \bar{e})/s$ so that their sample mean and sample variance are equal to 0 and 1, respectively. Then, the empirical cumulative distribution function $S_{z_n}(x)$ of the sample of $z_n$ values is computed. It is defined by \cite{42}

$$S_{z_n}(x) = \frac{1}{N} \times \text{ (number of } z_n \text{ s less than or equal to } x).$$

Similarly, the standard normal cumulative distribution function $S_{z_n}^0(x)$ is computed from the standard normal probability density function. Finally, the Lilliefors test statistic $T_L$ is defined by

$$T_L = \sup_x |S_{z_n}(x) - S_{z_n}^0(x)|$$

that is, the maximum distance of the two cumulative distribution functions.

**Decision rule.** The hypothesis $H_0$ is rejected at the significance level $\gamma$ if $T_L$ exceeds the $(1-\gamma)$ quantile as given in a table of quantiles of the Lilliefors test for normality as provided, for example, in Ref. [46]. MATLAB’s statistics toolbox [45] includes a function, called `lillietest.m`, for testing normality of a sample by means of the Lilliefors test.

### 5.2.4. Test for zero mean of the residuals

In addition to the Lilliefors test, which can be used to test the hypothesis that the residuals are normally distributed, a hypothesis testing method to test if the mean of the residuals is equal to 0, as proposed in Ref. [42], can be used. This method assumes the residuals to be independent, normally distributed with common but unknown variance. The null hypothesis $H_0$ and the alternative hypothesis $H_1$ are given by

$H_0$: the mean of the residuals is equal to 0

and

$H_1$: the mean of the residuals is not equal to 0.

**Test statistic.** It is known that for normally distributed residuals a $100(1-\alpha)$% confidence interval for the mean is given by

$$\left( \bar{e} - t_{N-1,1-\alpha/2} \frac{s}{\sqrt{N}}, \bar{e} + t_{N-1,1-\alpha/2} \frac{s}{\sqrt{N}} \right)$$

with $\bar{e}$ the sample mean as defined by Eq. (61), $s$ the standard deviation, that is, the square root of the sample variance $s^2$ as defined by Eq. (62), and $t_{N-1,1-\alpha/2}$ the $(1-\alpha/2)$ quantile of the $t$ distribution with $N-1$ degrees of freedom. The meaning of a $100(1-\alpha)$% confidence interval is that the probability that this interval covers the unknown true mean is $(1 - \alpha)$ [42]. This confidence interval can now be used to derive the following test for zero mean residuals.

**Decision rule.** A possible test using the confidence interval described above is given by: reject $H_0$ if the confidence interval (65) does not contain 0. Such a test has size $\alpha$ since the probability that $H_0$ is rejected if $H_0$ is true is $\alpha$. 

5.2.5. The lack of fit test

Obviously, assessment of the adequacy of the expectation model is easier if the experimenter has replications available (obtained from repeating the same experiment) [41]. A well known and widely used model assessment test is the so-called lack of fit test. The lack of fit test is based on the hypothesis $H_0$ to be tested is given by

$$H_0: w_{nj} = f_n(\tau) + v_{nj}$$

$(n = 1, \ldots, N; j = 1, \ldots, J_n)$

(66)

with $f_n(\tau)$ the proposed expectation model, $w_{nj}$ the observations, and the $v_{nj}$ are independent, identically normally distributed with zero mean and variance $\sigma^2$. The alternative hypothesis $H_1$ is that the observations are not described by Eq. (66). The lack of fit test is based on an analysis of variance (ANOVA) in which the residual sum of squares

$$S_r = \sum_{n=1}^{N} \sum_{j=1}^{J_n} (w_{nj} - \bar{w}_{nj})^2$$

(68)

with

$$\bar{w}_{nj} = \frac{1}{J_n} \sum_{j=1}^{J_n} w_{nj}$$

(69)

and the lack of fit sum of squares

$$S_f = S(\tilde{\tau}) - S_r = \sum_{n=1}^{N} J_n (\bar{w}_{n\bullet} - f_n(\tilde{\tau}))^2.$$  

(70)

The ML estimate $\tilde{\tau}_{ML}$ is obtained by maximizing the likelihood function derived from the joint probability density function of all $w_{nj}$. Since we consider normally distributed observations, $\tilde{\tau}_{ML}$ equals the least squares estimate. Notice that the ML estimate $\tilde{\tau}_{ML}$ is found by minimizing $\sum_{n=1}^{N} \sum_{j=1}^{J} (w_{nj} - f_n(\tilde{\tau}))^2$ with respect to $\tau$, which is equivalent to minimizing $\sum_{n=1}^{N} J_n (\bar{w}_{n\bullet} - f_n(\tilde{\tau}))^2$ with respect to $\tau$. Furthermore, notice that $S_r/(J - N)$ is an unbiased, model independent estimator of the noise variance $\sigma^2$. $S_r/\sigma^2$ will have a $\chi^2_{J-N}$ distribution, that is, a chi-square distribution with $J - N$ degrees of freedom.

**Test statistic.** The test statistic is given by [23]

$$T_{LF} = \frac{S_f}{S_r} \frac{J - N}{N - R}.$$  

(71)

If $H_0$ is true and the model $f_n(\tau)$ is a linear function of $\tau$, $S_f/\sigma^2$ has a $\chi^2_{N-R}$ distribution and $S_r$ and $S_f$ are independent. This means that for a linear model the test statistic $T_{LF}$ (being the ratio of two independent $\chi^2$ distributed random variables divided by their respective degrees of freedom) has Fisher’s $F$ distribution with $N - R$ and $J - N$ degrees of freedom if $H_0$ is true [41]. Because of asymptotic linearity, the lack of fit test is approximately valid for large $J$ when the model is nonlinear [23].

**Decision rule.** The range of $T_{LF}$ values for which the hypothesis $H_0$ is accepted at the significance level $\gamma$ is given by the interval

$$\left(F_{N-RJ-N,\gamma/2}, F_{N-RJ-N,1-\gamma/2}\right)$$

(72)

with $F_{N-RJ-N,\gamma/2}$ and $F_{N-RJ-N,1-\gamma/2}$ the $(\gamma/2)$ and $(1 - \gamma/2)$ quantile of Fisher’s $F$ distribution with $N - R$ and $J - N$ degrees of freedom, respectively [42].

6. Confidence regions and intervals for ML parameter estimates

Suppose that the expectation model used has not been rejected and that the ML estimates of the parameters have been computed. Then, in order to evaluate the level of confidence to be attached to these estimates, confidence regions and intervals associated with these estimates are
required. A 100(1 – \(z\)% confidence region is a region in \(R\)-dimensional parameter space that attempts to “cover” the true but unknown parameter vector \(\tau\) with probability \(1 - z\) [48]. A 100(1 – \(z\)% confidence interval for an element \(\tau_r\) of \(\tau\) is an interval that covers the true element \(\tau_r\) of \(\tau\) with probability \(1 - z\). Generally, confidence regions may be awkward to compute and difficult to represent graphically, especially if \(R\), the dimension of the parameter space, is larger than 3 (or even 2). Usually, one is more interested in confidence intervals. Confidence intervals are usually presented as the estimated value of the parameter plus or minus a certain amount (which is sometimes referred to as the measurement error or the error bar). Both regions and intervals will be considered in this section. Ideally, confidence regions and intervals are obtained by repeating the same experiment over and over again and subsequent evaluation of the statistics of the estimates obtained. However, this scenario is often not very realistic in practice. The methods that will be summarized in the present section do not require replications and are therefore more practically useful. In Section 6.1, likelihood confidence regions and intervals will be described. These regions and intervals are generally applicable in the sense that they can be derived for any distribution of the observations. The same applies to the confidence intervals described in Sections 6.2 and 6.3, which are based on the asymptotic normality of the Fisher score and the ML estimator, respectively. Section 6.4 describes confidence regions and intervals for parameter estimates obtained by applying the ML method to normally distributed observations, which corresponds to (weighted) least-squares estimation.

Notice that the present section aims to review existing methods and is largely based on Refs. [23,41,48,49].

6.1. Likelihood based confidence regions and intervals

A method to derive approximate confidence regions and intervals for the ML parameter estimates is the so-called likelihood method. Its applicability is invariant to the distribution of the observations. Let \(q(t)\) be the log-likelihood function of the parameters \(t\) of the expectation model \(f(t)\). Then, under certain regularity conditions, the hypothesis \(H_0 : \tau = \tau_0\) can be tested using the statistic

\[
LR = 2[q(\hat{\tau}_{ML}) - q(\tau_0)]
\]  

(73)

which is asymptotically \(\chi^2\) distributed when \(H_0\) is true. An asymptotically valid 100(1 – \(z\)% confidence region for \(\tau\) is then given by

\[
\{\tau : 2[q(\hat{\tau}_{ML}) - q(\tau)] \leq \chi^2_{R,1-z}\}
\]  

(74)

with \(\chi^2_{R,1-z}\) the \((1 - z)\) quantile of a chi-square distribution with \(R\) degrees of freedom. Under certain conditions, the above theory also holds if \(q(t)\) is replaced by a concentrated log-likelihood function (like Eq. (43)) [50].

An asymptotically valid 100(1 – \(z\)% confidence interval for the individual elements \(\tau_r\) of \(\tau\) is obtained by finding the values of \(\tau^*\) such that

\[
2[q(\hat{\tau}_{ML}) - q(\tau^*)] = \chi^2_{1,1-z}
\]  

(75)

in which \(\tau^*\) is the ML solution obtained by maximizing the log-likelihood function \(q(\tau)\) with respect to \(\tau\), given that the element \(\tau_r\) is held constant. The most extreme values of \(\tau_r\) that provide a solution to Eq. (75) represent the required limits of the confidence interval. This confidence interval is the projection onto the appropriate parameter axis of the confidence region (74). The method can be illustrated by means of the following example.

Example 9 (normally distributed observations).

For ML estimation from normally distributed observations with covariance matrix \(W\), the log-likelihood function is described by Eq. (36). Then it follows from Eqs. (36) and (74), that an asymptotically valid 100(1 – \(z\)% confidence region for \(\tau\) is given by

\[
\{\tau : [d^T(\tau)W^{-1}d(\tau) - d^T(\hat{\tau}_{ML})W^{-1}d(\hat{\tau}_{ML})] \leq \chi^2_{R,1-z}\}
\]  

(76)

with

\[
d(\tau) = w - f(t).
\]  

(77)

An asymptotically valid 100(1 – \(z\)% confidence interval for the confidence intervals for the individual elements \(\tau_r\) of \(\tau\) is obtained by finding
the values of $\tau^*$ such that

$$ [d^T(\tau^*)W^{-1}d(\tau^*) - d^T(\hat{\tau}_{ML})W^{-1}d(\hat{\tau}_{ML})] = \chi^2_{1,1-\alpha} $$

(78)

in which $\tau^*$ is the ML solution obtained by maximizing the log-likelihood $q(\tau)$ function with respect to $\tau$, given that the element $\tau_r$ of $\tau$ is held constant. The most extreme values of $\tau_r$ that provide a solution to (78) represent the required limits of the confidence interval. For uncorrelated observations having equal variance $\sigma^2$, Eqs. (76) and (78) simplify to

$$ \left\{ \tau : \sum_n d_n^2(\tau) - \sum_n d_n^2(\hat{\tau}_{ML}) \leq \sigma^2 \chi^2_{1,1-\alpha} \right\}, $$

(79)

and

$$ \sum_n d_n^2(\tau^*) - \sum_n d_n^2(\hat{\tau}_{ML}) = \sigma^2 \chi^2_{1,1-\alpha}, $$

(80)

respectively, with

$$ d_n(t) = w_n - f_n(t). $$

(81)

Notice that in order to construct Eqs. (79) and (80), $\sigma^2$ must be known. If $\sigma^2$ is unknown, it plays the role of a nuisance parameter that has to be estimated simultaneously with the elements of $\tau$. Approximate confidence regions and intervals can then be obtained by substituting the concentrated log-likelihood function (43) for $q(t)$ in Eqs. (74) and (75) yielding

$$ \left\{ \tau : N \left( \ln \sum_n d_n^2(\tau) - \ln \sum_n d_n^2(\hat{\tau}_{ML}) \right) \leq \chi^2_{1,1-\alpha} \right\}, $$

(82)

and

$$ N \left( \ln \sum_n d_n^2(\tau^*) - \ln \sum_n d_n^2(\hat{\tau}_{ML}) \right) = \chi^2_{1,1-\alpha}, $$

(83)

respectively. An alternative, more accurate likelihood based confidence region and interval for the special case of normally distributed observations with equal but unknown variance will be derived in Section 6.4.2.

Notice that likelihood based confidence regions and intervals may be disjoint and unbounded because the contours of the likelihood function may be disjoint and unbounded. Furthermore, the computation of likelihood based confidence regions and intervals is computationally expensive since it requires the evaluation of the log-likelihood function $q$ at a sufficient number of points to produce contours of constant likelihood. Nevertheless, using software packages such as MATLAB [45], the computation of likelihood based confidence regions and intervals is feasible indeed.

6.2. Confidence regions based on the asymptotic normality of the Fisher score

Alternatively, confidence regions can be obtained by using the fact that the Fisher score (7) is asymptotically (multivariate) normally distributed with a mean equal to the null vector and a covariance matrix equal to the information matrix $F$. This means for a large number of observations approximately:

$$ \frac{\partial q}{\partial \tau} \sim \mathcal{N}_R(0, F) $$

(84)

with $\mathcal{N}_R$ denoting the R-variate normal distribution and $F$ the Fisher information matrix. Then, it can be shown that, approximately,

$$ \left( \frac{\partial q}{\partial \tau} \right)^T F^{-1} \left( \frac{\partial q}{\partial \tau} \right) \sim \chi^2_R $$

(85)

so that an approximate 100(1 - $\alpha$)% confidence region for $\tau$ is given by

$$ \left\{ \tau : \left( \frac{\partial q}{\partial \tau} \right)^T F^{-1} \left( \frac{\partial q}{\partial \tau} \right) \leq \chi^2_{R,1-\alpha} \right\}. $$

(86)

Mostly, the CRLB $F^{-1}$ will be a function of the parameters to be estimated and is therefore unknown. Recall that the CRLB is given by

$$ F^{-1} = \left( -E \left[ \frac{\partial^2 q}{\partial \tau^2} \right] \right)^{-1}. $$

(87)

This expression can be rewritten as [23]

$$ F^{-1} = \left( -E \left[ \sum_n \left( \frac{\partial q}{\partial f_n(\tau)} \frac{\partial^2 f_n(\tau)}{\partial \tau^2} \right) \right] + J^T(\tau) \frac{\partial^2 q}{\partial \tau^2} J(\tau) \right)^{-1} $$

(88)
with \( J(\tau) \) defined by Eq. (13). Eq. (88) reduces to

\[
F^{-1} = \left( -\mathbb{E} \left[ J'(\tau) \frac{\partial^2 q}{\partial \tau^2} J(\tau) \right] \right)^{-1}
\]  

(89)

since \( \mathbb{E}[\partial q/\partial f_n(\tau)] = 0 \), because \( \partial q/\partial f_n(\tau) \) is the Fisher score of the parameter \( f_n(\tau) \) and, as mentioned in Section 3, a property of the Fisher score is that its expectation is equal to zero. In order to derive the confidence intervals (86), the following approximations of \( F^{-1} \) are commonly used \[49]:

\[
\hat{F}_a^{-1} = \left( -\sum_n \frac{\partial q(\hat{\tau}_{\text{ML}})}{\partial f_n} \frac{\partial^2 f_n(\hat{\tau}_{\text{ML}})}{\partial \tau^2} - J'(\hat{\tau}_{\text{ML}}) \frac{\partial^2 q(\hat{\tau}_{\text{ML}})}{\partial \tau^2} J(\hat{\tau}_{\text{ML}}) \right)^{-1}
\]  

(90)

and

\[
\hat{F}_b^{-1} = \left( -J'(\hat{\tau}_{\text{ML}}) \frac{\partial^2 q(\hat{\tau}_{\text{ML}})}{\partial \tau^2} J(\hat{\tau}_{\text{ML}}) \right)^{-1}
\]  

(91)

Alternatively, \( F^{-1} \) may be approximated by substituting \( \hat{\tau}_{\text{ML}} \) for \( \tau \) in the expression for the CRLB, which is given by (87), yielding

\[
\hat{F}_e^{-1} = \left( -\mathbb{E} \left[ \frac{\partial^2 q}{\partial \tau^2} \right] \right)^{-1}_{\tau = \hat{\tau}_{\text{ML}}}
\]  

(92)

Notice that approximations (90) and (91) differ only when

\[
\sum_n \frac{\partial q(\hat{\tau}_{\text{ML}})}{\partial f_n} \frac{\partial^2 f_n(\hat{\tau}_{\text{ML}})}{\partial \tau^2}
\]  

(93)

evaluated at \( \hat{\tau}_{\text{ML}} \), is nonzero. This term is a linear combination of the quantities \( \partial q(\hat{\tau}_{\text{ML}})/\partial f_n \), \( n = 1, \ldots, N \), with \( f_n \) the model \( f_n(\tau) \) evaluated at \( \hat{\tau}_{\text{ML}} \). It can be shown that the quantity \( \partial q/\partial f_n \) in Eq. (93) tends to a stochastic variable with an expectation equal to zero if \( f_n(\hat{\tau}_{\text{ML}}) \) tends to \( f_n(\tau) \), since then the quantity \( \partial q(\hat{\tau}_{\text{ML}})/\partial f_n \) is equal to the Fisher score of the parameter \( f_n \). Furthermore, notice that Eq. (93) will be equal to zero if the expectation model is linear in the parameters. If, additionally, the observations are normally distributed, the CRLB \( F^{-1} \) will be independent of \( \tau \) and the confidence region (86) becomes exact. As we will see below, for normally distributed observations approximations (91) and (92) are identical.

Notice that since the statistic \( (\partial q/\partial \tau)F^{-1}(\partial q/\partial \tau)^T \) depends only on the true parameter value \( \tau \) and not on the ML estimate \( \hat{\tau}_{\text{ML}} \) (approximate) confidence regions can also be found without calculating the ML estimate. Following this approach, an approximate \( 100(1-\alpha)\% \) confidence region is obtained from Eq. (86) by evaluating \( F \) and \( \partial q/\partial \tau \) at hypothesized values of \( \tau \). The confidence region is then formed by all values of \( \tau \) for which Eq. (86) is satisfied.

Finally, it is worthwhile mentioning that just like likelihood based confidence regions, confidence regions based on the asymptotic normality of the Fisher score may be disjoint and unbounded. Their computation is also computationally expensive (although not infeasible) since it requires in any case the evaluation of the gradient of the log-likelihood function \( q \) at a sufficient number of points to produce contours.

### 6.3. Confidence intervals based on the asymptotic normality of the ML estimator

Alternatively, confidence intervals can be obtained from the asymptotic normality of the ML estimator. Recall that the ML estimator of the parameter vector \( \tau = (\tau_1, \ldots, \tau_R)^T \) is known to be asymptotically multivariate normally distributed with a mean equal to the true value of the parameter vector and a covariance matrix equal to the CRLB. This means for a large number of observations approximately

\[
\hat{\tau}_{\text{ML}} \sim \mathcal{N}_R(\tau, F^{-1})
\]  

(94)

with \( \mathcal{N}_R \) denoting the \( R \)-variate normal distribution and \( F^{-1} \) the CRLB. Next, let us consider the function \( \rho(\tau) = a^T \tau \), with \( a \) a \( R \times 1 \) vector of constants. Then it follows from Eqs. (94) and (29) that, asymptotically:

\[
a^T \hat{\tau}_{\text{ML}} \sim \mathcal{N} (a^T \tau, a^T F^{-1} a).
\]  

(95)

Setting \( a^T = (0, 0 \ldots 0, 1, 0 \ldots 0) \) where the \( r \)th element of \( a \) is one and the remaining elements are zero, an approximate \( 100(1-\alpha)\% \) confidence interval for the \( r \)th element of \( \tau \), \( \tau_r \), is
given by

\[ \hat{\tau}_{ML, r} = \frac{1}{N - R} \sum_{n} d_{n}(\hat{\tau}_{ML}) \left( \frac{\partial^{2}f_{n}(\hat{\tau}_{ML})}{\partial t^{2}} \right) \]

in which the \( w_{mn} \) are the elements of \( W^{-1} \). Notice that (98) is also obtained by substituting the ML estimates \( \hat{\tau}_{ML} \) for the true parameters in the expression for the CRLB (22), which corresponds to Eq. (92). Substitution of Eq. (97) or Eq. (98) in Eq. (96) yields an approximate 100(1 − \( \alpha \))% confidence interval for \( \tau_{r} \). For uncorrelated observations having equal variance \( \sigma^{2} \), Eqs. (97) and (98) simplify to

\[ \hat{\tau}_{a}^{-1} = \sigma^{2} \left( \sum_{n} d_{n}(\hat{\tau}_{ML}) \right) \left( \frac{\partial^{2}f_{n}(\hat{\tau}_{ML})}{\partial t^{2}} \right) \]

+ \( J^{T}(\hat{\tau}_{ML})J(\hat{\tau}_{ML}) \)^{-1} \tag{99}

and

\[ \hat{\tau}_{b}^{-1} = \sigma^{2}(J^{T}(\hat{\tau}_{ML})J(\hat{\tau}_{ML}))^{-1}, \tag{100} \]

respectively.

Next, suppose that \( \sigma^{2} \) is unknown. Then the confidence intervals (96) cannot be computed. This problem can be circumvented by making use of the fact that, under certain regularity conditions, asymptotically,

\[ \sum_{n} d_{n}^{2}(\hat{\tau}_{ML}) \sim \chi^{2}_{N - R} \]

and, independently,

\[ \sqrt{\left[ F^{-1} \right]_{rr}} \sim \mathcal{N}(0, 1). \tag{102} \]

It is known that the ratio of a standard normally distributed random variable and the square root of an independently distributed chi-square random variable divided by its degrees of freedom has a Student’s \( t \)-distribution with the same number of degrees of freedom [42]. Hence, it follows from Eqs. (101), (102) and (25) that

\[ \frac{[\hat{\tau}_{ML}] - \tau_{r}}{s \sqrt{\left[ C^{-1} \right]_{rr}}} \sim t_{N - R} \]

with

\[ C = J^{T}(\tau)J(\tau) \]

and

\[ s = \sqrt{\frac{1}{N - R} \sum_{n} d_{n}^{2}(\hat{\tau}_{ML})}. \]

Therefore, an approximate 100(1 − \( \alpha \))% confidence interval for the \( r \)th element of \( \tau \), \( \tau_{r} \), is given by

\[ [\hat{\tau}_{ML}]_{r} \pm t_{N - R, 1 - \alpha/2} s \sqrt{\left[ C^{-1} \right]_{rr}} \]

with \( t_{N - R, 1 - \alpha/2} \) the \((1 - \alpha/2)\) quantile of a \( t \)-distribution with \( N - R \) degrees of freedom. For nonlinear models, the matrix \( C \), which equals \( F \sigma^{2} \), is a function of the true parameter vector \( \tau \), which is of course unknown. Approximations of \( C^{-1} \) are given by \( \hat{\tau}_{a}^{-1} / \sigma^{2} \) and \( \hat{\tau}_{b}^{-1} / \sigma^{2} \), where \( \hat{\tau}_{a}^{-1} \) and \( \hat{\tau}_{b}^{-1} \) are given by Eq. (99) and Eq. (100), respectively. The usefulness of (106) depends on the curvature
properties of $E[w] = f(\tau)$. The confidence intervals are exact if the parameters to be estimated enter the expectation model linearly. For nonlinear models, the intervals are only approximate.

**Example 11 (Poisson distributed observations).** For independent Poisson distributed observations, the log-likelihood function is given by

$$
q(t) = \sum_n [-f_n(t) + w_n \ln f_n(t) - \ln w_n!].
$$

Then it follows from Eqs. (16), (90), (91), (92) and (107) that

$$
\hat{F}_a^{-1} = \left(-\sum_n \left[-1 + \frac{w_n}{f_n(\hat{\tau}_{ML})}\right] \frac{\partial^2 f_n(\hat{\tau}_{ML})}{\partial t^2} + J^T(\hat{\tau}_{ML}) \text{diag}(w_n f_n^{-2}(\hat{\tau}_{ML}))^{-1} \right),
$$

and

$$
\hat{F}_b^{-1} = (J^T(\hat{\tau}_{ML}) \text{diag}(w_n f_n^{-2}(\hat{\tau}_{ML})))^{-1}.
$$

Substitution of Eq. (108), (109) or (110) in Eq. (96) then yields an approximate 100(1 - $\alpha$)% confidence interval for $\tau_r$.

It follows from Eq. (96) that unlike likelihood based and Fisher score based confidence intervals, the confidence intervals based on the asymptotic normality of the ML estimator will never be disjoint. In addition, they are much easier to compute.

6.4. Confidence regions and intervals based on normally distributed observations with unknown but common variance

It was shown in Section 4 that the ML estimator for independent, normally distributed observations with common variance corresponds to the uniformly weighted least-squares estimator. The least-squares estimator is well known and often used in practice. Its wide spread use is justified by the fact that practical observations are often well described by a normal distribution. This has to do with the fact that in practice, different error sources may simultaneously contribute to the observations. These contributions will generally be differently distributed. However, it follows from the central limit theorem that the sum of all these distributions will converge to a normal distribution. Many methods to find confidence regions and intervals for least-squares estimates can be found in the literature. This section aims to give a brief description of three leading methods for the special case of normally distributed observations with unknown but common variance: the lack of fit method, the likelihood method and the linearization method.

6.4.1. Lack of fit method

Let $I$ be the $N \times N$ identity matrix. Furthermore, let us define the $N \times N$ matrix $P(\tau)$ as

$$
P(\tau) = J(\tau)[J^T(\tau)J(\tau)]^{-1}J^T(\tau).
$$

The lack of fit method is then based on the fact that the quadratic forms

$$
Q_1(\tau) = d^T(\tau)P(\tau)d(\tau)/\sigma^2
$$

and

$$
Q_2(\tau) = d^T(\tau)(I - P(\tau))d(\tau)/\sigma^2
$$

with $d(\tau)$ defined by Eq. (77), are independent chi-squared random variables with $R$ and $N - R$ degrees of freedom, respectively [23]. Since the ratio of two independent chi-squared random variables divided by their respective degrees of freedom is distributed as Fisher’s $F$-distribution, the ratio of $Q_1(\tau)/R$ and $Q_2(\tau)/(N - R)$ is distributed as $F_{R,N-R}$. A 100(1 - $\alpha$)% confidence region for $\tau$ is thus given by [48]

$$
\left\{ \tau : \frac{d^T(\tau)P(\tau)d(\tau)/R}{d^T(\tau)(I - P(\tau))d(\tau)/(N - R)} \leq F_{R,N-R,1-\alpha} \right\}
$$

(114)
with \( F_{R,N-R,1-z} \) the \((1-z)\) quantile of the \( F \) distribution with \( R \) and \( N-R \) degrees of freedom. The thus obtained confidence region is exact, but computationally expensive, since both \( f(\tau) \) and \( P(\tau) \) have to be evaluated at a sufficient number of points to produce the required contours.

Similarly, a lack of fit confidence interval for the \( r \)th element \( \tau_r \) of the parameter vector \( \tau \) can be derived (see, e.g., Ref. [49]). However, the thus obtained interval is only approximate. In addition, it is even more computationally expensive than the lack of fit confidence region and thus unlikely to be used routinely [49]. Moreover, to the authors’ knowledge, there exists no empirical evidence that the lack of fit intervals are superior to those obtained by the so-called likelihood method, which is less complicated to compute and will be presented in the next subsection.

### 6.4.2. Likelihood method

Recall that the problem of finding confidence regions and intervals for ML estimates from normally distributed observations with unknown but equal variance was already considered in Example 9, where a method based on the concentrated log-likelihood function was discussed. In this section, we will review an alternative likelihood method that is based on the fact that the quadratic forms

\[
\sum_n d_n^2(\tilde{\tau}_{\text{ML}}) \quad \text{(115)}
\]

and

\[
\sum_n d_n^2(\tau) - \sum_n d_n^2(\tilde{\tau}_{\text{ML}}) \quad \text{(116)}
\]

with \( d_n(t) = w_n - f_n(t) \) are (approximately) independent chi-squared distributed random variables with \( N-R \) and \( R \) degrees of freedom, respectively. (Notice that it was shown in example 5 that \( \tilde{\tau}_{\text{ML}} \) can be obtained without knowledge of \( \sigma^2 \).) Hence, it can be shown that

\[
\frac{\sum_n d_n^2(\tau) - \sum_n d_n^2(\tilde{\tau}_{\text{ML}})}{\sum_n d_n^2(\tilde{\tau}_{\text{ML}})} \cdot \frac{N-R}{R} \sim F_{R,N-R} \quad \text{(117)}
\]

since the ratio of two independent \( \chi^2 \) distributed random variables divided by their respective degrees of freedom is distributed as Fisher’s \( F \)-distribution. Therefore, a \( 100(1-z)\% \) confidence region for \( \tau \) is given by

\[
\left\{ \tau : \sum_n d_n^2(\tau) - \sum_n d_n^2(\tilde{\tau}_{\text{ML}}) \leq R s^2 F_{R,N-R,1-z} \right\} \quad \text{(118)}
\]

with \( s \) defined by Eq. (105) and \( F_{R,N-R,1-z} \) the \((1-z)\) quantile of the \( F_{R,N-R} \) distribution. According to Ref. [49], the limits of the likelihood method \( 100(1-z)\% \) confidence interval for \( \tau_r \) can be obtained by finding the points that maximize \( (\tau_r - [\tilde{\tau}_{\text{ML}}])^2 \) subject to

\[
\sum_n d_n^2(\tau) - \sum_n d_n^2(\tilde{\tau}_{\text{ML}}) \leq s^2 F_{1,N-R,1-z} = s^2(i_{N-R,1-z/2})^2. \quad \text{(119)}
\]

This confidence interval is the projection on to the appropriate parameter axis of the confidence region given by Eq. (118).

**Bates and Watts’ curvature measures.** It can be shown that the validity of the likelihood confidence regions and intervals depends on the degree of nonlinearity in the model. The confidence regions and intervals derived are exact if the parameters to be estimated enter the expectation model linearly. For nonlinear models, they are only approximate. The usefulness of Eqs. (118) and (119) thus depends on the nonlinearity (i.e., curvature) properties of \( \mathbb{E}[w] = f(\tau) \). Bates and Watts have proposed curvature measures for normally distributed observations that can be used as diagnostic tools to predict whether or not the confidence regions and intervals produced by the likelihood methods will be reliable [51]. A Monte Carlo study has indicated the high effectiveness of these tools [49]. Bates and Watts show that two types of curvature can be distinguished: parameter-effects curvature and intrinsic curvature [51]. The derivation of Bates and Watts’ curvature measures is straightforward, but outside the scope of this paper. With respect to Eq. (118), it is important to notice that the term \( \sum_n d_n^2(\tilde{\tau}_{\text{ML}}) \) is invariant to reparameterizations. This means that we can assume that a transformation has been found to remove the parameter-effects curvature [23]. Hence, Eq. (118) is only affected by intrinsic curvature, which is often negligible [23]. Finally, it
should be noted that, generally, the degree of nonlinearity, and therefore the inadequacy of confidence regions such as Eqs. (118) and (119) decreases with the number of observations.

Notice that the (approximate) likelihood regions and intervals are generally somewhat easier to compute than the exact confidence regions produced by the lack of fit method described in the previous subsection. However, they are still computationally expensive compared with the linearization method described in the following section.

6.4.3. Linearization method

Linearization methods for constructing confidence regions and intervals assume that the expectation model \( f_n(\tau) \) can be adequately approximated by a linear (or affine) approximation to the model at the ML estimate \( \hat{\tau}_{\text{ML}} \). Linearization methods are based on the fact that if the model \( f(\tau) \) is linear in the parameters \( \tau \), an exact 100(1 - \( \alpha \))% confidence region for \( \tau \) is given by

\[
\{ \tau : (\tau - \hat{\tau}_{\text{ML}})^T C^{-1} (\tau - \hat{\tau}_{\text{ML}}) \leq s^2 R F_{N-R,1-\alpha} \}
\]

(120)

with \( C \) defined by Eq. (104), whereas an exact 100(1 - \( \alpha \))% confidence interval for \( \tau \) is given by

\[
[\hat{\tau}_{\text{ML}}, \hat{\tau}_{\text{ML}} + t_{N-R,1-\alpha/2} s \sqrt{C^{-1} \tau}] .
\]

(121)

It can be shown that for linear models the confidence region (120) corresponds to Eq. (118). The confidence interval (121) corresponds to the interval defined by Eq. (106).

If the model \( f_n(\tau) \) is linear in the parameters \( \tau \), the Jacobian matrix \( J \) and thus the matrix \( C \) does not depend on the parameters. Under the assumption that \( f_n(\tau) \) can be adequately linearized, approximate confidence regions and intervals can be obtained by substitution of

\[
(J^T (\hat{\tau}_{\text{ML}})) J (\hat{\tau}_{\text{ML}}))^{-1}
\]

(122)

for \( C^{-1} \) in Eqs. (120) and (121), respectively. Notice that Eq. (122) corresponds to \( \hat{F}_b / \sigma^2 \), with \( \hat{F}_b \) described by Eq. (100). Other choices of \( C^{-1} \) have been suggested, but it has been reported that Eq. (122) is to be preferred [49]. Although the linearized confidence regions and intervals appear to be the most approximate for nonlinear models, they are relatively easy to compute. Therefore, the linearization method is the most commonly used method and it has been implemented in several software packages such as MATLAB [45] and EXPHER [52]. Unlike the lack of fit and the likelihood-based region, which may have irregular shapes, the linearization region is always ellipsoidal [48]. The confidence regions and intervals derived in this section are exact for linear models. For nonlinear models, they can be used as a linearized approximation. One must always be aware of the fact that the extent to which the linearized confidence region and interval approximate the exact confidence region and interval depends on the extent of nonlinearity of the model. As with the likelihood confidence regions and intervals, the Bates and Watts curvature measures may be used as diagnostic tools to predict whether the confidence regions and intervals produced by the linearization methods will be reliable or not [51].

7. Conclusions

In this paper, it has been shown how the maximum likelihood (ML) method can be used to estimate structure parameters including those of electron microscopy images. The use of the ML method is motivated by the fact that the ML estimator has favorable statistical properties, such as consistency, asymptotic efficiency and asymptotic normality. It has been shown that the application of the ML method requires the joint probability density function of the observations (image pixel values) and its dependence on the structure parameters to be estimated to be known. This dependence is usually established by the availability of an expectation model, which is a (usually physics based) model describing the expectation values of the observations. A thus parameterized joint probability density function can be used for two purposes. First, it can be used to derive an expression for the highest precision with which the structure parameters can be estimated without bias. Second, it can be used to derive the ML estimator of these parameters. It
has been made clear that there is no use in attaching confidence to (ML) parameter estimates as long as the model has not been accepted. Several model assessment techniques have been discussed. Finally, it has been shown how to construct confidence intervals for ML parameter estimates. In part two of this two-part paper, the theory reviewed in this paper will be applied to real-world HRTEM images.

Acknowledgements

The research of Dr. A.J. den Dekker has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences. Dr. S. Van Aert gratefully acknowledges the financial support of the Fund for Scientific Research—Flanders (FWO)

References